Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 2 (2010) 446–453

Procedia Environmental Sciences

International Society for Environmental Information Sciences 2010 Annual Conference (ISEIS)

Water Simulation Method Based on BPNN Response and Analytic Geometry

Zhiming Zhang, Xiaoyan Wang*, Yang Ou

College of Resources, Environment & Tourism, Capital Normal University, Beijing 100048, China

Abstract

With three-year monthly monitoring data (from Jan. 2007 to Dec. 2009) of ten given sections of one river in Beijing, a BP neural network of DO and principle components are built by taking the concentration of DO as the evaluation criterion of the water quality. Three orthogonality principle component vectors are exacted through correlation analysis and principal component analysis with other monitoring parameters (water temperature, pH, COD_{Cr}, COD_{Mn}, BOD, NH_3-N). After the evaluation on the availability of the net with root mean square error (RMSE) and the average relative error of the net's response of the test set, the 3-D figure is built by the net's response and the three PC vectors. For the continuous output response of the BPNN, the errors of the untrained samples closed to the trained ones are trivial. The status of water quality and the location in the 3-D space are represented by a sample selected with the cluster analysis. As the center of sphere, the value of PC vectors is used to find the shortest vector for raising the DO concentration inside the sphere. According the contribution of the monitoring parameters to each PC vector, the most sensitive parameter to the DO concentration can be found to meet the demand of the required DO standard. This method finds the empirical function between DO and temperature, BOD, NH_3-N, which can be used to make up for the missing data and reduce the monitorial indexes, also could serve as a reference for the water quality management.

Keywords: water quality simulation; principal component (PC); back-propagation neural network (BPNN); vector

* Corresponding author. E-mail: cnuwxy@sohu.com

1878-0296 © 2010 Published by Elsevier Open access under CC BY-NC-ND license.
doi:10.1016/j.proenv.2010.10.049

1. Introduction

The dissolved oxygen (DO) is an important quality index of some water. But it is difficult to simulate the DO concentration by the traditional mathematical method due to the different effect factors on the different waters [1].

The traditional models based on the mechanism work only in the simplified conditions. For example, Shao Xiongfei predicted the pollutant's distribution and the diffusion in Qiantang River with 2-D model. But the grid size was designed too big to ensure the precision[2]; Duan Dehong simplified the river section to be calculated into a straight one without considering the influence of the curved section over the water flow and the pollutant distribution when conducting the water quality simulation of the multi-sewage river [3].

The artificial neural network is an empirical model which has nothing to do with the mechanism, only input and output matters. Instead of analyzing the mechanism in the actual process, it carries out the statistics analysis according to the materials acquired from the actual process and then concludes a mathematical expression between the parameters and the variables based on the minimum error principle. In the field of the water resource management, it has been widely used for water quality prediction, monitoring sites optimization, water amount need prediction, flow's variation and water quality evaluation. Therefore, this article simulates the DO concentration with the BP artificial neural network, and then offers the reference for the water quality improvement based on the simulation result.

2. BPNN Establishment

2.1. Source of Data

The data used here is the monthly data of a total of ten sections of Beiyun River from Jan.2007 to Dec.2009, with seven parameters including water temperature T(℃), pH, DO(mg/l), COD_{Cr} (mg/l), COD_{Mn}(mg/l), BOD(mg/l), NH_3-N(mg/l).

2.2. Data Optimization

2.2.1. Parameter Selection

To reduce the unnecessary information as well as lessen the neurons in the input layer, calculation of the coefficients correlation with DO can be carried out to confirm the parameters of the BPNN[4].

Table 1 The coefficient of correlation between DO and other water quality parameters

	DO	T	pH	COD_{Cr}	COD_{Mn}	BOD	NH_3-N
DO	1.00	-0.84	-0.11	-0.07	-0.12	0.29	-0.43

This study takes the coefficients of a considerable modulus related closely with the DO concentration such as water temperature, BOD, and NH3-N as the used variables of the network.

2.2.2. Normalization

The existence of the different units will result in the erroneous contribution of the parameters, especially those with large value. This article conducts the normalization with the following equation[5]:

$$PN = \frac{2 \times (P - \min p)}{(\max p - \min p)} - 1 \quad (1)$$

In this equation, *P* and *PN* represent the input data before and after the change, separately, max p and min p the maximum value and the minimum value that the premnmx function gets.

2.3. BPNN Establishment and Simulation

2.3.1. Neural Network Training Samples Selection

In establishing the neural network model, whether the training samples are representative [6] and whether the amount of the samples is adequate will have a great impact on the simulation outcome. Therefore, after averaging the values of the monthly monitorial data of the sections from Jan.2007 to Dec.2009, this article conducts a CA of the normalized data (The parameters involved for CA are DO, BOD, NH3-N).

The result of the cluster analysis is as follows:

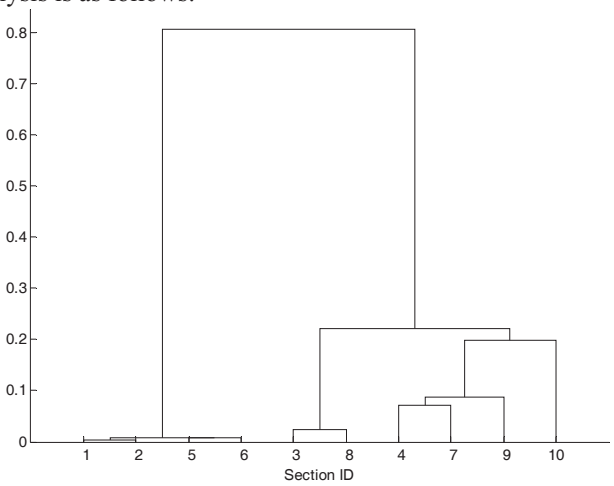

Fig.1 The result of CA with three-year average data on the different sections

Section 1 is very similar to Section 2. Considering the result of the cluster analysis and the different landuse patterns around sections, Section 1 and Section 2 are thought as the same type, using the monthly monitorial data of three years (seventy-two samples) of the two sections as the training set (47 data) and the test set (25 data) of the neural network. The DO concentration of the two sections changes as follows:

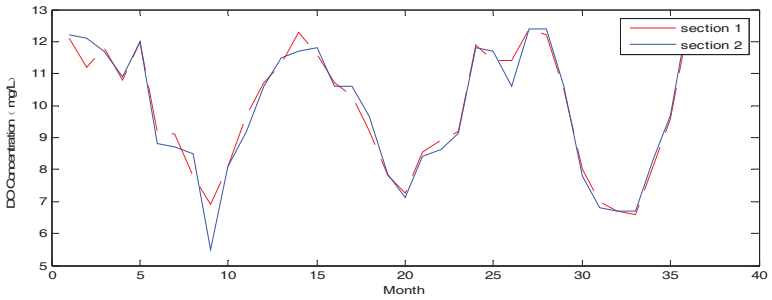

Fig.2 The Comparison of monthly DO Concentration at different sections of same type

The purpose of the orthogonal transformation of the normalized data by means of Principal Component Analysis (PCA) is to reduce the correlation of the neurons in the input layer of ANN[7]. The correlation of the dimensions can be reduced to the lowest after transforming the data from the original index sphere to the principle component sphere. The transformation matrix is:

$$T = \begin{pmatrix} -0.6871 & 0.1330 & -0.7143 \\ 0.5755 & -0.5005 & -0.6467 \\ -0.4435 & -0.8554 & 0.2674 \end{pmatrix}$$

2.3.2. ANN Training and Simulation

Because of the limited parameters in this article, it is unnecessary to reduce the dimensions by means of PCA. But after the orthogonal transformation, the neurons in the input layer of the neural network can be regarded as the independent variables. Three-layer BP neural network with trainlm is selected as the training function with the upper limit of the training times being 5000 and the number of the neurons in the hidden layer 7. The transmitting function from the input layer to the hidden layer is logsig and hidden layer to the output one is tansig. After training, the training and simulation ability of the neural network is presented in the following pictures.

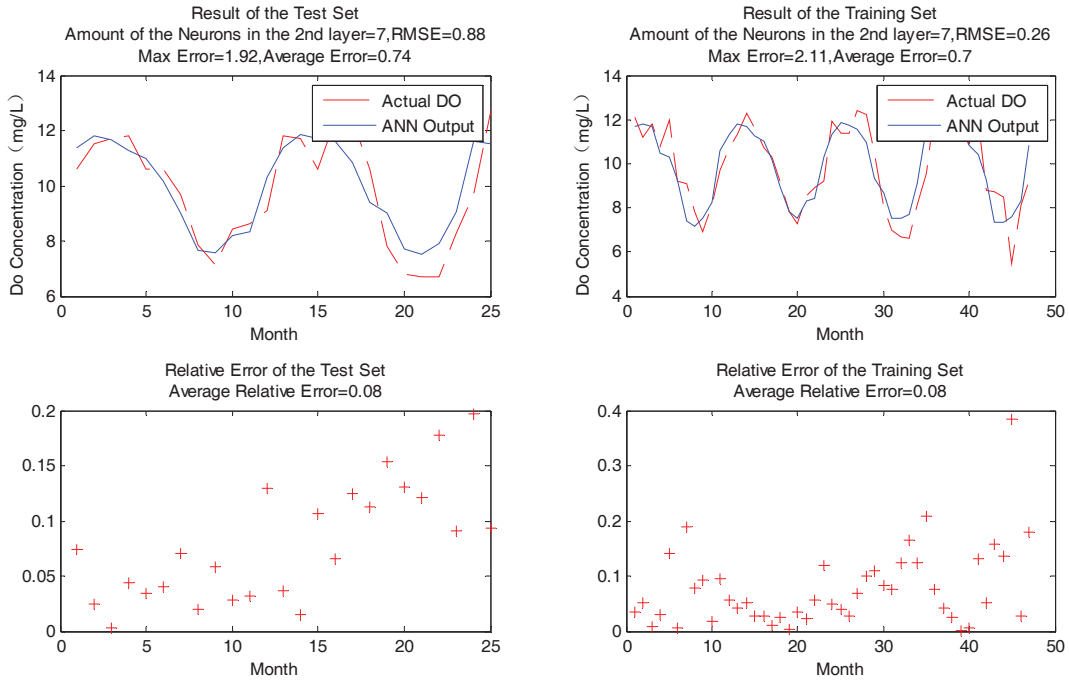

Fig.3 The result of simulation of test set and training set

In the figure, the left part is the result of the test while the right part is the result of the training. It can be seen that the DO concentration of the trained neural network can be calculated through three orthogonal principle component variables. And the average relative error is no more than 8% with a good generalization of the network.

3. Response Result Analysis Based on ANN

3.1. Weight Analysis

Considering that the contribution degree of the input variables to the output ones of the neural network can be expressed by the relatively important index I, the calculating formula is as follows [4] [8] [9]:

$$I = \frac{\sum_{j=1}^{S_1} |w_{ij}|}{\sum_{i=1}^{r} \sum_{j=1}^{S_1} |w_{ij}|} \tag{2}$$

In this formula, w_{ij} represents the connecting weight between the neuron i in the input layer and the neuron j in the hidden layer. r、S_1 represent the number of the neurons in the input layer and those in the hidden layer,

separately.

After calculation, the result is as follows:

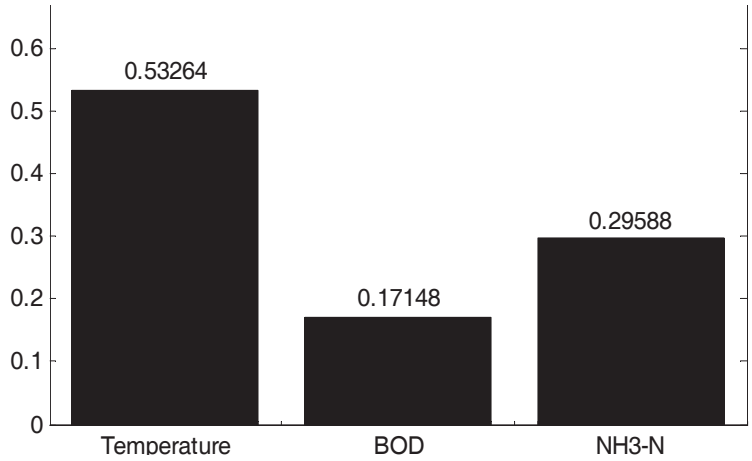

Fig.4 The influence of the three parameter selected in ANN

From the result of the neural network response, of the three parameters, temperature has a biggest influence over the DO concentration, NH_3-N medium and BOD the least. But this method operates according to the weight value from the input layer to the hidden one without considering the transfer function, threshold and the function effect from the hidden layer to the output one. Therefore, this relatively important index can only serve as a reference with poor accuracy. Consequently, this article offers a method about factor effect analysis based on neural network response.

3.2. Factor Effect Analysis Based on Neural Network Response

3.2.1. Principle

Because temperature, BOD and NH_3-N are all continuous variables, the principle component variables as their linear combinations are also continuous. The following picture is a slice of the neural network response value distribution of three principle component variables. According to the gradual changing effect of the colors in the picture, it is easy to see that the response curve of the BP network is continuous. So, the network responses of which points are closed to the precise-simulated point are also precise enough.

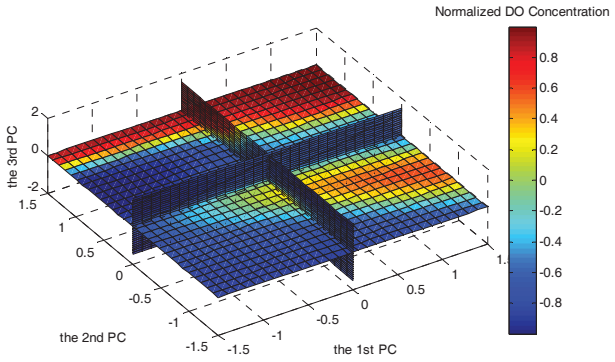

Fig.5 Slice of response of BPNN

Because of the orthogonality of the principle component vectors in the space, a sphere can be drawn with r (r value is decided with the result of ANN generalization) as its radius and the point that can reflect the water quality characteristic as its center. Because of the generalization of BP neural network, the coordinates of all the points within the sphere can get a precise DO concentration through the trained neural network function effect. And the changing rate of the DO concentration varies at the different directions. By decomposing the vector of the direction where the DO concentration grows the fastest at each orthogonal direction, the changing amount of other indexes can be confirmed at the time of increasing the DO concentration of the section at the fastest speed according to the anti-transformation of the orthogonal transformation.

3.2.2. Realization

3.2.2.1. Representative Data Selection

The representative data should be a sample of a high prediction precision. And the data of the training set has a higher precision than that of a test set. Therefore, it should be the training set where to find the representative samples. After a Euclidean distance CA of the data from the training set (47 samples), the outcome is as follows:

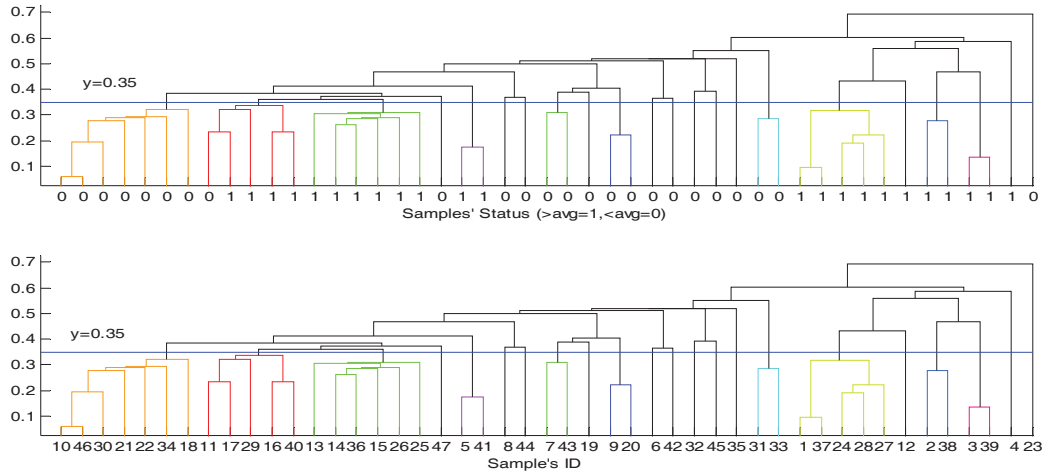

Fig.6 The result of CA of the monthly data from the two selected sections

It can be seen that through CA, the part of higher DO concentration can be parted from that of lower DO concentration and each part can be further divided into several parts (Compared to the Average DO concentration). Therefore, in choice of the representative samples, the criteria that the samples could represent should be taken into consideration. This article only discusses the orange part (the sample numbers are 10, 18, 21, 22, 30, 34, 46) and these samples represent the section with its lowest DO concentration.

After calculating the distances between the seven samples, No.46 sample that has the smallest total of distances to other samples is chosen as a representative. The data is as follows:

Table2 The parameters of the representative sample

DO(mg/l)	Temperature(℃)	BOD(mg/l)	NH_3-N(mg/l)
8.1	22.5	2	0.065

After a principle component transformation, the coordinate of the representative sample in the space built by the principle component vectors is (-0.0781, 0.4120, -0.0723).

3.2.2.2. Response Space Calculation

Each principle component vector is equally divided into thirty pieces within [-1.5, 1.5], to record the neural network response value of each box. Supposing r=0.1, find the point of the highest DO concentration within the sphere with the representative sample as its center and 0.1 as its radius, and then establish the vector V', from the sphere centre to the point of which the neural network output DO concentration is C, and the DO concentration of the sphere centre is C_0. With $V=T^T \times V'$, the following result can be acquired:

$$V= (0.0437, -0.0327, -0.0378),\ C= 8.55,\ C0=8.10$$

This shows that as the DO concentration increases along its fastest way, the temperature rises while BOD and NH_3-N falls accordingly. However, the influence extent is similar, which is much different from the outcome of the calculation of the relatively important indexes. But the result is the same when arranging them from the most to the least by influence, or Temperature>NH_3-N>BOD. It shows the water quality management in the similar sections, must pay attention to NH_3-N and BOD input both at the same time.

3.3. Discussion

Three parameters out of a total of seven are selected with the correlation analysis, and the data are extended by cluster analysis, an ANN as a water quality model is built. The influence of each factor with the response are discussed.

As a water quality model, it can simulate DO concentration by three parameters such as temperature, BOD, NH_3-N with an average relative error no more than 8%. It is an effective way to find index with correlation. The other parameters may not be used into DO simulation to reduce the redundancy information.

Water quality management through neural network response has its advantage in the exclusion of the human factors, using a model built by the data on hand to indicate the relationship among the water quality parameters, while in a neural network response space, it's difficult to determine the response accuracy far away from the original data points. Besides, the response value near some samples of an unpleasant fitting effect could probably produce a substantial error.

Therefore, when it comes to a representative sample selection, the prediction accuracy is preferred as well as the representability. And a distance range (r) should be set to limit the radius of the hyper-sphere in predicting the water quality of some visual point. Otherwise, the reliability of the model will be greatly affected. Consequently, the reliability drops as r value increases in spite of a better water quality optimization.

If the costs of the variations of each index are considered, the result of the model should change. But now they are not considered yet.

4. Conclusion

Under the premise of the similarity of two sections, this article selects the data of the two sections for ANN establishment. The average relative error is no more than 8% .

With the cluster analysis of the three-year data of the selected sections, this article finds that the DO concentration in the water has a direct relationship with the typical combinations of the three variables of temperature, BOD, NH_3-N.

Based on the result of ANN, this article offers a method of factors influence analysis, and draws a conclusion that to DO concentration in selected sections, temperature contributes the most then NH_3-N, and BOD the least, but varies little.

References

[1] Chen Lihua, Ma Shengquan, Li Li. A model to evaluate do of river based on artificial neural network and style book. Journal of Hainan Normal University (Natural Science). 2008;21 (4):372-376

[2] Shao Xiongfei. Application of mathematical simulation to the water quality prediction in the project of drawing water from Qiantang River to Hangzhou City . Environmental Pollution & Control, 2005;27(6):465-467

[3] Wang Xu,Li Kefeng,Li Ran,Hao Hongsheng,Tuo Youcai. Water quality simulation of coupled effect of multiple sewage outfalls in urban river reach . Engineering Journal of Wuhan University, 2008;41(2):24-27

[4] Xu Min,Li Yunsheng,Zeng Guangming,LiangJie. Modeling Changes of Non-point Source Pollution Load for Watershed Using Bayesian Regularized BP Neural Network. Environmental Science & Technology, 2009; 32(11):171-176

[5] Zhou Kaili, Kang Yaohong. Neural Network Models and Emulation Programming Design Tsinghua Press.

[6]Sovan Lek,Guegan J F． Artificial neural networks as a tool in ecological modeling, an introduction． Ecological Modeling, 1999;120:65-73．

[7] Long Xunjian, QianJu, Liang Chuan. Water demand forecast model of BP neutral networks based on principle component analysis . Journal of Chengdu University of Technology (Science & Technology Edition), 2010;37 (2):206-210

[8]Joseph H W Lee, Yan Huang, Mike Dickman, A.W.Jayawardena. Neural network modelling of coastal algal blooms . Ecological Modeling, 2003;159: 179~201

[9]Pei Hongping, Luo Nina, Jiang Yong. Applications of back propagation neural network for predicting the concentration of chlorophyll-a in West Lake. Acta Ecologica Sinica, 2004;24(2):246-251